Aalto University
School of Electrical
Engineering

Speech Recognition System Improvements for North Sámi Speaker Dependent and Speaker Independent Tasks

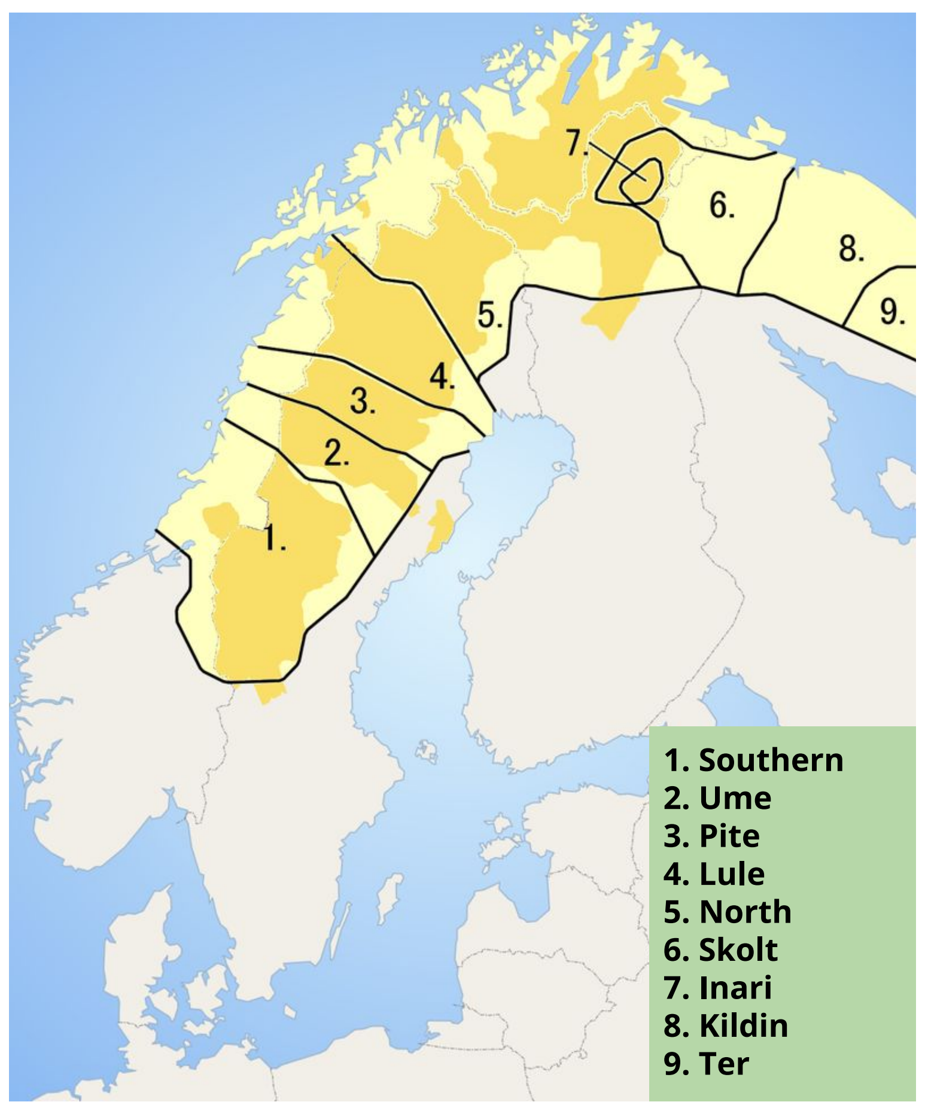
Aku Rouhe, Mikko Kurimo
Department of Information and Communications Engineering, Aalto University, Finland

Context

- Previously we built North Sámi speech recognisers using wav2vec 2.0 and tested HMM/DNN vs. Attention-based models - HMM/DNN was better, but still had high WER. [1]
- Here we improve the HMM/DNN to see how far we can get with our current data.

North Sámi

- Sámi languages are spoken in areas which fall inside Northen Norway, Sweden, Finland, and Russia. They are in the Uralic language family.
- North Sámi (*Davvisámegiella*) is the biggest of the Sámi languages
- DATA:
  - UIT-SME TTS Corpus: 7.6h, 2 speakers
  - Giellagas North: 1.6h, 19 speakers
  - Freecorpus: >700k lines of text
- No official test sets or data splits, so we created our own.

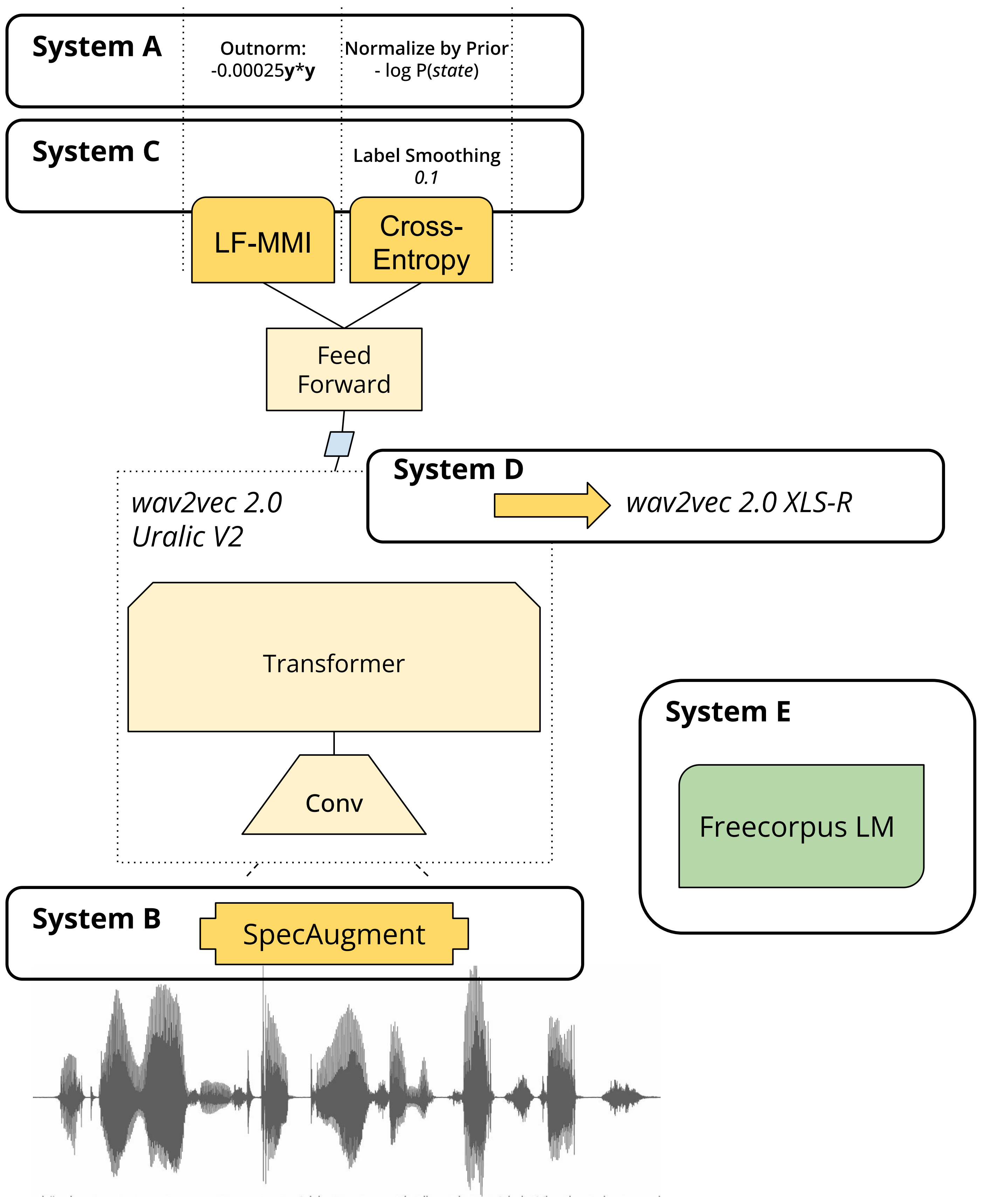


1. Southern
2. Ume
3. Pite
4. Lule
5. North
6. Skolt
7. Inari
8. Kildin
9. Ter

ASR System Improvements

The baseline recipe has a Kaldi-style HMM/GMM, wav2vec 2.0 Uralic V2 Large (300M params. 41kh Finnish, Hungarian, Estonian) AM body, using two output heads (even in decoding). The baseline language model uses the training data transcripts.

Cumulative improvements (Systems A-E):



System A
Outnorm: -0.00025**y*y**
Normalize by Prior - log P(*state*)

System C
Label Smoothing *0.1*

LF-MMI
Cross-Entropy
Feed Forward

wav2vec 2.0 Uralic V2

System D
wav2vec 2.0 XLS-R

Transformer
Conv

System E
Freecorpus LM

System B
SpecAugment

North Sámi Results

	WER/CER [%]	
	Valid	Test
Speaker Independent		
Baseline	79.20 / 56.76	72.66 / 46.53
System A	78.54 / 49.84	71.36 / 40.00 †
System B	80.22 / 52.36	71.15 / 41.50 †
System C	77.80 / 51.40	70.52 / 41.93 †
System D	77.52 / 49.37	69.09 / 38.93 †
System E	**76.21** / **48.90**	**66.43** / **38.63** †
Speaker Dependent		
Baseline	45.85 / 25.21	51.78 / 29.65
System A	47.80 / 22.27	52.02 / 25.99
System B	50.73 / 22.18	51.95 / 24.25
System C	47.48 / 22.08	50.85 / 24.71 †
System D	**45.69** / 21.82	50.51 / 24.76 †
System E	46.18 / **20.91**	**47.85** / **22.54** †

†: Cumulative improvements shown to be significant by bootstrap estimate [2]

Speaker Independent
Train: 8.0h, 7 speakers
Valid: 0.2h, 4 speakers
Test: 1.5h, 10 speakers
No overlap in speakers

Speaker Dependent
Train: 9.1h, 21 speakers
Valid: 0.1h, 11 speakers
Test: 0.5h, 11 speakers
Valid and Test speakers appear in Train

Finnish Results

We validated our findings in a Finnish experiment

	WER/CER [%]	
	Valid	Test
Speaker Independent		
Baseline	13.93 / 4.93	10.13 / 3.34
System A	13.43 / 4.56	10.02 / 3.14
System B	13.29 / **4.45**	**9.65** / 3.05
System C	**13.14** / 4.47	9.66 / **3.04**
System D	18.48 / 5.25	15.09 / 3.72

Speaker Independent
Train: 20h, 340 speakers
Valid: 2.8h, 10 speakers
Test: 2.8h, 10 speakers
No overlap in speakers. Subsets from Finnish Parliament Train16 data.

Takeaways

- Successful changes (A-C) from targeting model overconfidence
- XLS-R (400kh speech, 128 languages) better than Uralic V2 for North Sámi but not Finnish
- Unsuccessful experiments:
  - Taking output from non-final layer of wav2vec 2.0
  - Tying acoustic states more heavily
  - Training with an initial phase where wav2vec 2.0 is frozen
  - Decreasing subword vocabulary size
- North Sámi needs a public benchmark test set.

References

[1] A. Rouhe, A. Virkkunen, J. Leinonen, and M. Kurimo, "Low Resource Comparison of Attention-based and Hybrid ASR Exploiting wav2vec 2.0," in *Proc. Interspeech 2022*, 2022, pp. 3543–3547.

[2] M. Bisani and H. Ney, "Bootstrap estimates for confidence intervals in asr performance evaluation," in *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 2004, pp. I–409.

Read the paper:

Contact: Aku Rouhe
Email: aku.rouhe@aalto.fi